arXiv:1101.0566v1 [physics.bio-ph] 3 Jan 2011

Low energy pathways for reproducible *in vivo* protein folding

L. Cruzeiro*

CCMAR and FCT, Universidade of Algarve,

Campus de Gambelas, 8005-139 Faro, Portugal

(Dated: January 4, 2011)

Abstract

Two proteins, one belonging to the mainly α class and the other belonging to the α/β class, are selected to test a kinetic mechanism for protein folding. Targeted molecular dynamics is applied to generate folding pathways for those two proteins, starting from two well defined initial conformations: a fully extended and a α-helical conformation. The results show that for both proteins the α-helical initial conformation provides overall lower energy pathways to the native state. For the α/β protein, 30 % (40%) of the pathways from an initial α-helix (fully extended) structure lead to unentangled native folds, a success rate that can be increased to 85 % by the introduction of a well-defined intermediate structure. These results open up a new direction in which to look for a solution to the protein folding problem, as detailed at the end.

INTRODUCTION.

Proteins are mega molecules. Even a small protein, with 60 amino acids, possesses approximately one thousand atoms which are linked by covalent bonds, hydrogen bonds, electrostatic interactions and van der Waals interactions. Despite the inherent complexity of such a system, proteins, in cells, are capable of folding reproducibly to a well defined three-dimensional structure that is generally irregular and lacks symmetry. How do proteins do it? According the thermodynamic hypothesis [1] the native structure of proteins corresponds to the minimum of the free energy and protein folding is a thermodynamic process simply driven by thermal agitation. The complementary proposal that the free energy landscape is funnel-shaped was later put forward to explain also why folding can proceed comparatively fast [2–4]. However, in spite of many decades of study, and of the improvements in the accuracy of force fields [5–7] and in the power of computer facilities [8, 9], it has not yet been possible to determine a protein structure solely from its amino acid sequence and the thermodynamic and funnel hypotheses remain to be proven.

Here another possibility is considered, namely, that protein folding is a kinetic process and that the native structure is just one of the many kinetic traps in which proteins can find themselves in [10]. In previous works [11–14] it has been shown that a protein from a given CATH [15] class can be forced into artificial, non-native structures that belong to other CATH classes and maintain these structures for at least 50 nanoseconds. These results favour instead a *multi-funnel-shaped* free energy landscape of proteins and a specific *kinetic* process for *in vivo* protein folding was suggested [14]. For a kinetic process to lead to a well-defined three dimensional structure everytime, it must be associated with a specific pathway, as Levinthal first proposed [16]; however, and equally importantly, *it must always start from the same well-defined initial conformation*. In [14] it was suggested that this specific conformation, that is, the conformation that *all* proteins have immediately after synthesis, is helical and that the first step in folding is the bending of this initial helix at specific amino acid sites. The purpose of the present study is to make a preliminary test of the efficiency of such a kinetic process for folding. To that end, two proteins, representative of the mainly α and α/β CATH [15] protein classes, were selected and pathways from the initial conformation to the native state were generated using Targeted Molecular Dynamics (TMD) [17]. In TMD simulations, harmonic restraints are added to the protein force field in

order to drive an initial protein conformation to a given target conformation [17, 18]. TMD simulations were first applied to the T $\leftrightarrow$ R transition of the protein insulin [17] and have since been used to study a variety of other problems, including, e.g., the conformational changes associated with the functioning of a molecular motor [19], the elucidation of the reaction steps in the full catalytic cycle of a protein involved in an electron transfer process [20], as well as in protein folding [21]. Thus, TMD simulations have generally shown their usefulness to model processes that involve large conformational changes.

Experimentally, it has not yet been possible to determine the structure of nascent chains, that is, of the structure proteins have as they come out of the ribosomal tunnel but it is known that both fully extended and α-helical structures fit into the tunnel dimensions [22]. In the present study, TMD simulations are used to compare the efficiency of folding from an initial helical conformation, proposed previously to be the conformation of proteins immediately after synthesis [14] to the efficiency of folding from a fully extended conformation.

Two proteins, each representative of one of the four main protein CATH classes [15], were chosen and their native structures were obtained from the Protein Data Bank (PDB) [23]. One of these proteins (PDB1BDD [24], 60 amino acids, 941 atoms, mainly α) has a native structure constituted by three α-helices, while the second protein (PDB1IGD [25], 61 amino acids, 927 atoms, α/β) has a native structure that includes one α-helix and a β-sheet. Two initial conformations, both of which have been identified as viable experimentally [22], were taken for each of the two proteins, namely, one conformation in which the backbone is fully extended and a second conformation in which the backbone is folded into an ideal α-helix. The energy minimized versions of these two conformations were used as initial conditions for the TMD simulations. The ff99SB force field [26], with an implicit solvent, implemented in AMBER 9 [27], which has been shown to give a better representation of secondary structure [28], was used in all simulations. For each initial conformation, 20 independent TMD trajectories to the native state were generated by changing the seeds of the random forces in the Langevin thermal baths, with $T = 298$ K in all cases. In TMD simulations the harmonic forces that drive the initial conformations to the native structure arise from the following term added to the atomic potential energy function:

$$U_{\mathrm{TMD}} = \frac{1}{2}\, k\, N\, \left(\mathrm{RMSD}_N(t) - \mathrm{RMSD}_{\mathrm{target}}(t)\right)^2 \tag{1}$$

where N is the number of atoms used to calculate the root mean square deviation per

TABLE I. $\mathrm{RMSD}_{\mathrm{target}}(t = 0)$, in Å, for the two initial conformations of the two proteins.

PROTEIN	INITIALLY EXTENDED	INITIALLY α-helix
mainly α (1BDD)	55.78	21.54
α/β (1IGD)	59.76	24.46

atom (RMSD), $\mathrm{RMSD}_N(t)$ being the RMSD, with respect to the native structure, of the conformation the protein has at time t, and $\mathrm{RMSD}_{\mathrm{target}}(t)$ being the target RMSD at that time. In the simulations reported here, $k = 100$ kcal/mol/Å^2 and N is the total number of carbons, nitrogens and oxygens in the backbone of the two proteins selected.

In TMD simulations, $\mathrm{RMSD}_{\mathrm{target}}(t)$ is a linear function of time, being completely defined by the values at two time instants. Here, these two times are 1) its initial value, $\mathrm{RMSD}_{\mathrm{target}}(t = 0)$, given in table I, and 2) its final value, set to 0.1 Å in all trajectories. As shown in table I, for both proteins, the α-helical conformation is more than 2.4 times closer to the native structure than the fully extended conformation. Preliminary simulations starting from the α-helical conformation showed that the total potential energy (including the TMD term) does not vary much until $\mathrm{RMSD}_N \approx 6$ Å and so, in order to make the simulations with the two different initial conformations as equivalent as possible, in all the TMD simulations presented here, two $\mathrm{RMSD}_{\mathrm{target}}(t)$ functions were used, i.e., the first 0.1 ns were spent in the convergence of the initial conformation to within 6 Å of the native structure, that is, $\mathrm{RMSD}_{\mathrm{target}}(t = 0.1 \text{ ns}) = 6$ Å, and 0.4 ns were allowed for the further convergence to within 0.1 Å of the native conformation. In this way, a slower rate of change of the RMSD is imposed in the final stages of convergence to the native state, allowing more time for the side chains to avoid steric overlaps, and, most importantly, making the rate of RMSD change in this latter stage the same for all simulations. The overall duration of each TMD simulation was 0.5 ns, comparable to other TMD protein folding simulations [21].

Each of the two initial conformations was taken as an initial condition in 20 independent TMD simulations, as detailed above. Figure 1 shows the average over the 20 trajectories and the corresponding standard deviations of the total potential energy as a function of instantaneous value of RMSD_N for the mainly α 1BDD protein [24]. The curve in red is for the trajectories starting with a α-helical conformation and the curve in green is for the trajectories starting with a fully extended backbone conformation. Included in the values

displayed in figure 1 is the contribution of the TMD term (1) which, however, only starts to rise above 0.3 % of the total value when the RMSD distance to the native structure becomes less than 0.6 Å (not shown). This means that the energetics of these folding pathways is determined essentially by the atomic interactions in the ff99SB AMBER potential [26].

Figure 1 shows that the pathways from the initially extended conformations are populated by protein conformations which have, on average, a potential energy that is much greater than those that arise when the initial conformation is α-helical; indeed, even in the final stages of approach to the native state, the former protein conformations are at least 40 kcal/mol above the latter. Furthermore, contrary to what was found in previous folding simulations [21] and below for the α/β 1IGD protein, all pathways of the mainly α 1BDD protein lead to conformations with a perfectly folded backbone at least within 0.3 Å of the native structure.

To test further the hypotheses put forward in [14], another protein, the α/β 1IGD protein [25], whose native structure includes both an α-helix and a four stranded β-sheet, was used. As reported previously [21], many of the TMD folding simulations of the α/β 1IGD protein lead to final structures that, although apparently close to the native structure, differ from it by entangled backbone folds such as those highlighted by the red circles in figure 2. Indeed, inspection of the trajectories with the Visual Molecular Dynamics (VMD) software [30] reveals that 12 (14) of the 20 TMD simulations starting with a fully extended (α-helical) conformation lead to such entangled final structures. These entanglements can be resolved by letting the chains go through each other, an artificial process enabled by the TMD term (1) but of course penalized by extremely large values of the potential energy. Thus, in figure 3 the averages are made over viable pathways only, that is, over the TMD trajectories in which such backbone entanglements did not occur. Although the TMD term contributes a little more than before to the energetics of the pathways, its total amount only rises above 1 % when the RMSD distance to the native structure goes below 1 Å, with the larger values being associated with entangled structures (not shown). Figure 3 shows that, also for the α/β 1IGD protein, the viable pathways from an initial α-helix to the native state are much less energetic than the corresponding ones from a fully extended structure, with the conformations covered by the latter having a potential energy more than 130 kcal/mol greater than the former to start with, and with their values only merging when the RMSD distance to the native structure is approximately 6 Å.

In [14] it is suggested that the initial conformation of a protein is a helix and that the first step in protein folding is the bending of this helix at specific amino acid sites. In the case of proteins whose secondary structure is just a set of helix bundles, as in the mainly α 1BDD protein [24], this first step is the only major step in their folding pathways. VMD [30] animations of the TMD trajectories of the 1BDD protein do show pathways of this sort when the initial conformation is α-helical (not shown). (When the initial conformation is fully extended, on the other hand, the formation of the α-helices constitutes the last step, coming after the backbone has folded to the native topology). For proteins whose native structure includes both helices and sheets, as is the case of the α/β 1IGD protein, it was further suggested [14] that the formation of the β-sheets is due to destabilizing interactions between the amino acid side chains that are thus thrown together by the first step. VMD [30] animations of the TMD trajectories of the α/β 1IGD protein starting from the α-helical conformation reveal pathways that start off in that manner but in which the helical portions very quickly become distorted because the convergence to the right secondary structure is mixed with the convergence to the right backbone fold (not shown). This is to be expected because of the unspecific character of the TMD term (1). In order to test, in a more direct way, the effect that an intermediate all-helical conformation has on the folding efficiency of the α/β 1IGD protein, such a putative intermediate was generated by substituting the β-sheets in the native structure of I1IGD by α-helices, as shown in figure 4. This intermediate shall here be called an *embrio* because, according to the kinetic process proposed in [14], its structure is one of the main determinants of the final, native, structure of the protein.

Further TMD simulations were run by first driving an initial α-helix to the embrio, and then by driving the embrionic conformations to the native structure of 1IGD. As the RMSD between the α-helix and the embrio is 23.78 Å and the RMSD between the embrio and the native 1IGD structure is 15.05 Å, in order to keep approximately the same rate of change of RMSD as before, the first simulations had a duration of 0.05 ns and the second simulations had a duration of 0.45 ns. In the first 0.05 ns all 20 trajectories converge to within 0.1 Å of the embrio structure; furthermore, 17 out of the 20 trajectories from the embrio to the native state lead to un-entangled final structures (not shown). Figure 3, in which the potential energies of the protein conformations in these 17 viable folding pathways are displayed in blue, shows that the α-helix $\rightarrow$ embrio $\rightarrow$ native trajectories provide, on average, the lowest energy folding pathways for the α/β 1IGD protein.

This study does not prove the hypothesis that the α-helix is the conformation that all proteins have as they come out of the ribosomal tunnel; it does not prove either the hypothesis that the first step in the folding of all proteins is the formation of a compact core, here called an embrio, constituted only by helices (and disordered helices or turns), nor does it prove that β-sheets arise when one or more of the helices in that embrio is not stable; but it does show that those hypotheses are viable in that they can lead to low energy folding pathways to correctly folded, unentangled, native states. The simulations reported here also show that, of the two nascent chains that have been identified as experimentally feasible [22], that in the form of a helix is to be preferred to a fully extended conformation. Furthermore, the results here provide a glimpse of the many entangled states that can be expected to arise if protein folding is just a random search of the native structure, driven by thermal noise, especially if the initial structure is also arbitrary. Indeed, only 30 % and 40 % of the direct pathways to the native state, from a α-helical and a fully extended initial conformation, respectively, lead to a correctly folded 1IGD protein, even in the presence of the artificial TMD terms (1). On the other hand, the simulations here also show how a well defined intermediate/transient structure in the folding pathway can dramatically increase the odds of reaching the native structure in a reproducible manner: with the roughly built embrio tentatively considered here (see figure 4) the probability of reaching the native structure became 85 %! *The really important point, though, is the new direction in which to look for a solution to the protein folding problem that is opened up by these and other [11–14] results: instead of a thermodynamic process, a kinetic process; instead of arbitrary initial structures, a well defined and already ordered nascent chain; instead of many pathways, or even preferred pathways, a specific pathway for folding.* Another important point is the concrete program of research that naturally follows from it [14]: 1) the identification of the key amino acids, or amino acid sequences, at which the bends of the initial α-helix occur, can be made by a statistical analysis of the known structures of proteins formed by helical bundles only; 2) the same structures can be mined to find the relative orientations of the helical pieces in their respective embrios, and 3) once the key amino acids and these initial orientations have been identified, a statistical analysis of the known structures of α/β proteins may tell us which amino acid sequences in these initial helical pieces generate repulsive side chain interactions that, in turn, destabilize those helices and lead to the formation of β-sheets. It is indeed hoped that the results here will inspire further work to explore this new direction.

* Most of these simulations were performed at the Milipeia cluster of the Laboratory for Advanced Computing of the University of Coimbra, Portugal.

[1] C.B. Anfinsen, Science 181, 223 (1973).

[2] J.D. Bryngelson, J.N. Onuchic, N.D. Socci and P.G. Wolynes, Proteins 21, 167 (1995).

[3] J.N. Onuchic, Z. Luthey-Schulten and P.G. Wolynes, Ann. Rev. Phys. Chem. 48, 545 (1997).

[4] K.A. Dill, and H.-S. Chan, Nature Struct. Biol. 4, 10 (1997).

[5] M. Christen et al., J Computat. Chem 26, 1719 (2005).

[6] D.A. Case et al. J. Computat. Chem. 26, 1668 (2005).

[7] B. R. Brooks et al., J. Computat. Chem. 30, 1545 (2009).

[8] F. Allen et al., IBM Systems Journal 40, 310 (2001).

[9] See the web site: http://folding.stanford.edu/.

[10] L. Cruzeiro-Hansson and P.A.S. Silva, J. Biol. Phys. 27, S6 (2001).

[11] L. Cruzeiro, http://arxiv.org/abs/0712.2034.

[12] L. Cruzeiro, J. Phys. Org. Chem. 21, 549 (2008).

[13] L. Cruzeiro and P.A. Lopes, Mol. Phys. 107, 1485 (2009).

[14] L. Cruzeiro, "Protein Folding", in Specialist Periodical Reports Chemical Modelling, edited by M. Springborg, Applications and Theory Vol. 7, 89 (Royal Society of Chemistry, London, 2010).

[15] C.A. Orengo et al., Structure 5, 1093 (1997).

[16] C. Levinthal, J. Chim. Phys. 65, 44 (1968).

[17] J. Schlitter, M. Engels, P. Krüger, E. Jacoby, and A. Wollmer, Mol Simul 10, 291 (1994).

[18] J. Apostolakis, P. Ferrara, and A. Caflisch, J Chem Phys 110, 2099 (1999).

[19] V. Ovchinnikov, B. L. Trout, and M. Karplus, J. Mol. Biol. 395, 815 (2010).

[20] A. Negri, D. Rodríguez-Larrea, E. Marco, A. Jiménez-Ruiz, J. M. Sánchez-Ruiz, and F. Gago, Proteins: Struc. Funct. Bioinf. 78, 36 (2010).

[21] J. Apostolakis, P. Ferrara, and A. Caflisch, Proteins: Struct, Funct. and Gen. 39, 252 (2000).

[22] J.P. Ellis, P.H. Culviner, and S. Cavagnero, Prot. Sci. 18, 2003 (2009).

[23] H.M. Berman et al., Nucleic Acids Research 28, 235 (2000).

[24] H. Gouda et al., Biochemistry 31, 9665 (1992).

[25] T. Gallagher, P. Alexander, P. Bryan and G.L. Gillilan, Biochemistry, 33, 4721 (1994).

[26] V. Hornak et al. Proteins 65, 712 (2006)

[27] D.A. Case et al., (2006), AMBER 9, University of California, San Francisco.

[28] L. Wickstrom, A. Okur, and C. Simmerling, Biophys. J.

[29] L. Cruzeiro, J. Phys.: Condens. Matter 17, 7833 (2005).

[30] W. Humphrey, A. Dalke and K. Schulten, J. Molec. Graphics 14, 33 (1996).

CAPTION LIST

1. (color online) Average total potential energy (kcal/mol) over 20 independent TMD simulations, together with the corresponding standard deviations, as a function of the RMSD distance to the native structure of the mainly α 1BDD protein [24]. The starting conformation is fully extended (α-helical) for the green (red) curves. The inset shows the final stages of convergence to the native structure.

2. (color online) The native fold of the α/β 1IGD protein is depicted on the left and on the right the red circles highlight two of the most common backbone entanglements that arose in the TMD simulations of this protein. The figures were made with the VMD software [30].

3. (color online) Average total potential energy (kcal/mol) over non-entangled TMD simulations, together with the corresponding standard deviations, as a function of the RMSD distance to the native structure of the α/β 1IGD protein [25]. The starting conformation is fully extended (α-helical) for the green (red) curves and the embrio (see text and figure 4) for the blue curves. The inset shows the final stages of convergence to the native structure.

4. (color online) The putative embrio (see text) built from the native fold of the α/β 1IGD protein by substituting the native β-sheets by α-helices. This figure was made with the software VMD [30].

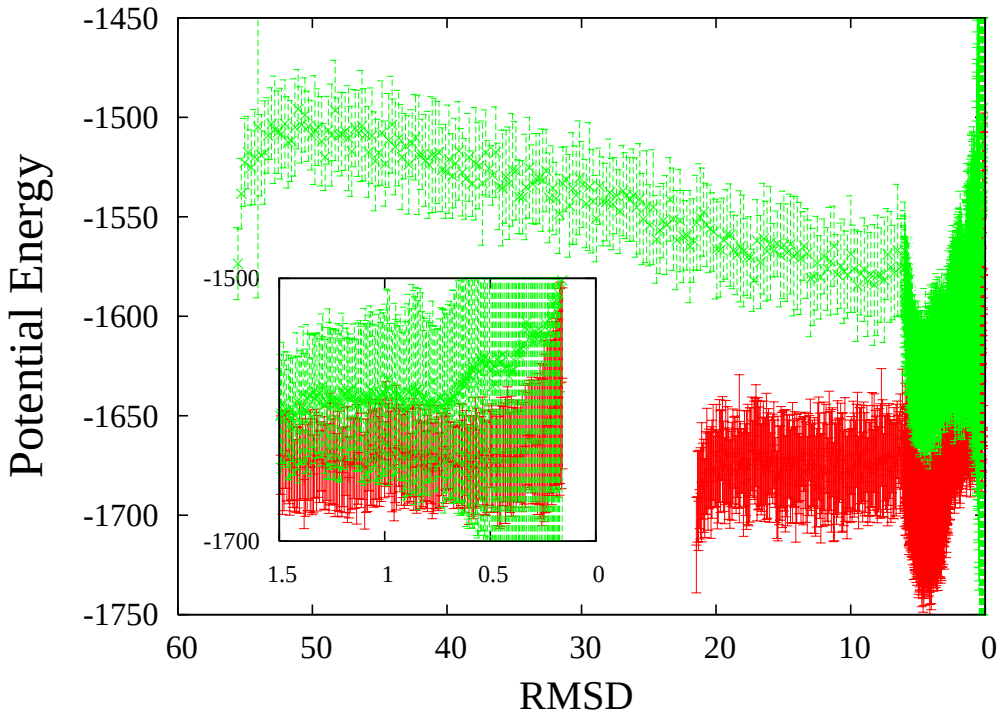
Potential Energy
RMSD
-1450
-1500
-1550
-1600
-1650
-1700
-1750
60
50
40
30
20
10
0
-1500
-1700
1.5
1
0.5
0

FIG. 1.

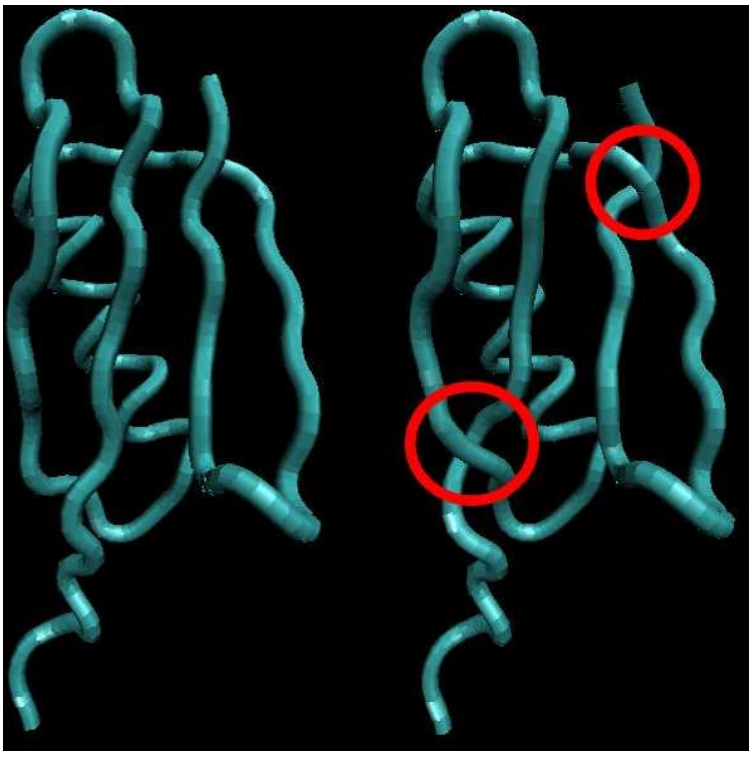

FIG. 2.

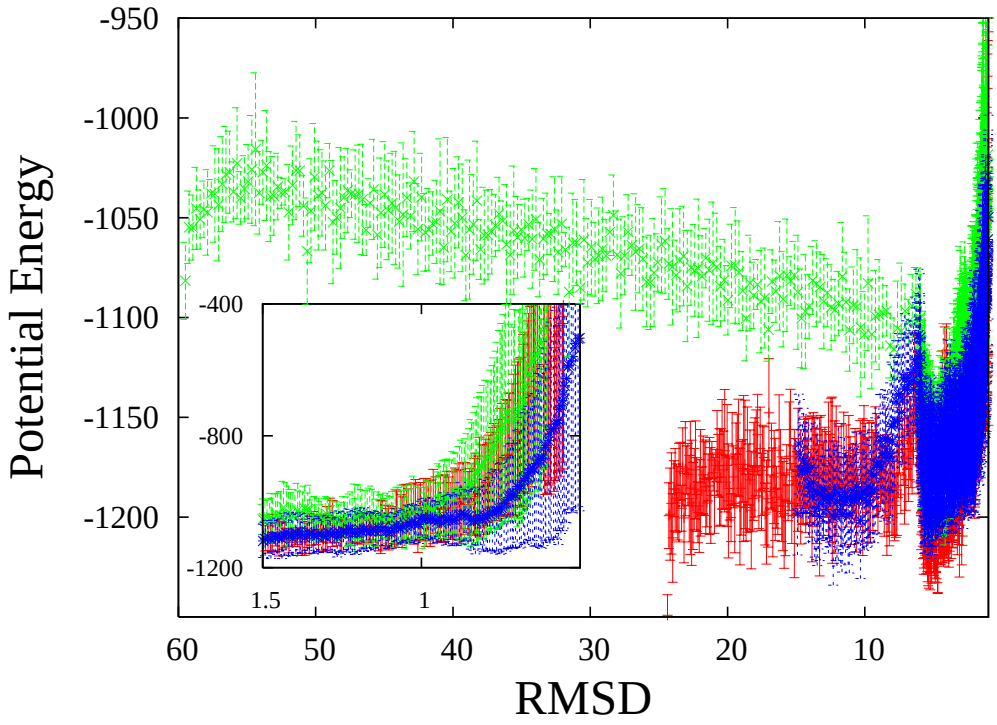
Potential Energy
-950
-1000
-1050
-1100
-1150
-1200
60
50
40
30
20
10
RMSD
-400
-800
-1200
1.5
1

FIG. 3.

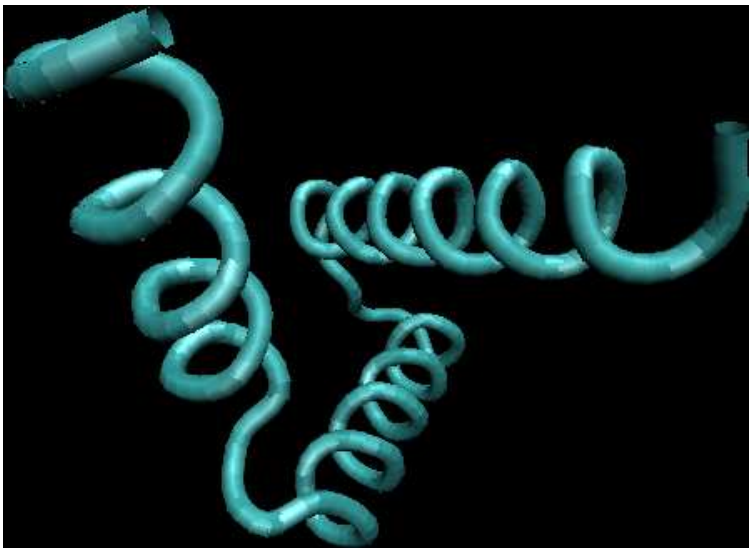

FIG. 4.